RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment

Year: 2023 | Citations: 619 | Authors: Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao

---

Abstract

Generative foundation models are susceptible to implicit biases that can arise from extensive unsupervised training data. Such biases can produce suboptimal samples, skewed outcomes, and unfairness, with potentially serious consequences. Consequently, aligning these models with human ethics and preferences is an essential step toward ensuring their responsible and effective deployment in real-world applications. Prior research has primarily employed Reinforcement Learning from Human Feedback (RLHF) to address this problem, where generative models are fine-tuned with RL algorithms guided by a human-feedback-informed reward model. However, the inefficiencies and instabilities associated with RL algorithms frequently present substantial obstacles to the successful alignment, necessitating the development of a more robust and streamlined approach. To this end, we introduce a new framework, Reward rAnked FineTuning (RAFT), designed to align generative models effectively. Utilizing a reward model and a sufficient number of samples, our approach selects the high-quality samples, discarding those that exhibit undesired behavior, and subsequently enhancing the model by fine-tuning on these filtered samples. Our studies show that RAFT can effectively improve the model performance in both reward learning and other automated metrics in both large language models and diffusion models.